Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 92 (2016) 520 – 525

Procedia Computer Science

2nd International Conference on Intelligent Computing, Communication & Convergence

(ICCC-2016)

Srikanta Patnaik, Editor in Chief

Conference Organized by Interscience Institute of Management and Technology

Bhubaneswar, Odisha, India

Performance Evaluation of Different Similarity Functions and Classification Methods using Web Based Hindi Language Question Answering System

Rajni Devi[a], Mohit Dua[b,*]

[a]*Department of Computer Engineering, Banasthali Vidyapith, Banasthali, Rajasthan, India*
[b]*Department of Computer Engineering, National Institute of Technology, Kurukshetra, Haryana, India*

Abstract

Question Answering (QA) system is an approach to extract the correct answer for the query asked by the user in its own language. The work discussed is implemented for Hindi Language objective type questions and answers. The paper implements the comparison of nine different similarity functions and two classification methods used to retrieve the desired information. The results revealconclude that Smith Waterman outperforms the other similarity functions in perforamnce evaluation. The K-Nearest Neighbor(K-NN) algorithm gives 97% , 95.6% and Neasret Neighbor (NN) algorithm gives 93.3%,95% for two differtent test data sets, respectively.

Peer-review under responsibility of the Organizing Committee of ICCC 2016

Keywords: Hindi Question Answering System, Machine Learning, Data Mining, Similarity functions, Text Similarity measure, classification method, Nearest Neighbor, K-Nearest Neighbor.

Peer-review under responsibility of the Organizing Committee of ICCC 2016
doi:10.1016/j.procs.2016.07.377

1. Introduction

In a Question Answering system user inputs its query in a natural language and in response the system gives its output answer in the same language. The vital application is the utility of natural language interface to database (NLIDB) system [1]. The developed system uses some objective type questions in Hindi language and verifies the corresponding answers. User can pick a question from a given database and the system gives answers from the database. It also evaluates that the given answer is accurate or not. The system gives the accurate answer instead of giving a number of results. Classification method plays an important role for getting accurate answer. For example, if a user enters a question "हिमाचलप्रदेशकीराजधानीक्याहै" and the system detects the answer correctly. The search range is reduced due to classification method and by using the knowledge based intelligent system [2]. In the given system the user deals with the different types of questions like क्या (what), कब (when), कौन (who), कहाँ (where) etc [3].

2. Related Work

The reviewed literature reveals that the QA system was born in 1950[4]. The first developed QA system was BASEBALL, built in 1961[5]. BASEBALL system used to answer questions about the particular game. In 1972 LUNAR was built for Apollo Moon Mission [5]. In 1972, systems were built for understanding the dialogue; SHRDLU and GUS [6]. SHRDLU was built for toy domain and to simulate robot while on the other hand GUS accessed the information about airline flights from a restricted database. In 1980's and 1990's the research about knowledge base system initiated and the system like MYCIN, used in medical, was built. The first web based QA system got developed between 2004 and 2006 (Frgetal) that introduced clustering method [3]. During 2009 to 2011 the system was implemented for Chinese language that used semantic web technology [7].

3. Architecture of QA System

The architecture of the developed system has been divided into two modules backend and frontend. As described by figure 1, the backend comprises of training functions whereas frontend performs the functions of question processing, feature extraction and application of classification methods.

3.1. Backend

3.1.1 Keyword Extraction: This step is same as tokenization as will be described in frontend. After extracting the keyword from the dataset those keywords are maintained in a table and corresponding to this table an index table is maintained. Firstly, a token number is provided to each token, for example: [1. भारत, 2. भूटान, 3. राजधानी, 4. क्या, 5. मनाया, 6. दिवस, 7. कब, 8. शहीद, 9. अंतर्राष्ट्रीय……], and then a VSM (Vector Space Matrix) is used. A three dimensional array is taken [QF, AF, QW], where QF refers to question field, AF is answer field, QW is considered as question word and each token number is stored in matrix corresponding to that word.

3.1.2 Index Maintainer: From this table the answer can be easily obtained via index number. Corresponding to the label an answer is obtained which is already stored in answer index table. The resulting answer is stored in a hashing table and hence, each index number has a different answer table related to the type of question. The N number of tables are maintained corresponding to index values.

3.1.3Train data: After performing the above said steps trained data set is obtained and user can easily acquire the answer from the trained data set.

3.2. Frontend

This is the second module of the implemented architecture where user inputs the query and get the related answer from the backend. The module comprises of:

3.2.1 Question Processing

In this sub module Hindi question entered by the user are first read and tokenization is applied on the same. In tokenization method the question is echeloned and Context words are extracted and these words are called Unigram features or Bag of Words. Any n successive words in a question are consider as a feature [8]. The repetition of words in question approach load value and this load value shows the consequence of word in a question [8]. For example: "राष्ट्रीय युवा दिवस कब मनाया जाता है" has Unigram representation as ({राष्ट्रीय,1} {युवा,1} {दिवस,1}{मनाया,1} {कब,1} {जाता,1} {है, 1}), Bigram representation as ({ राष्ट्रीययुवा,1}), Trigram representation as ({ राष्ट्रीययुवादिवस,1}), WH-words ({ क्या , कब , कहाँ etc.}) [8]. The purpose of preprocessing phase is to reduce the text. Stop words are removed and these are also called non-bearing words like की, का, केetc. For example: In "भारतकीराजधानीक्याहै" 'की' is a stop word. Removal of these types of words doesn't effect on structural type query and after removal of these word remaining word is "भारत" "राजधानी" "क्या" and then these all are mapped into different tables.

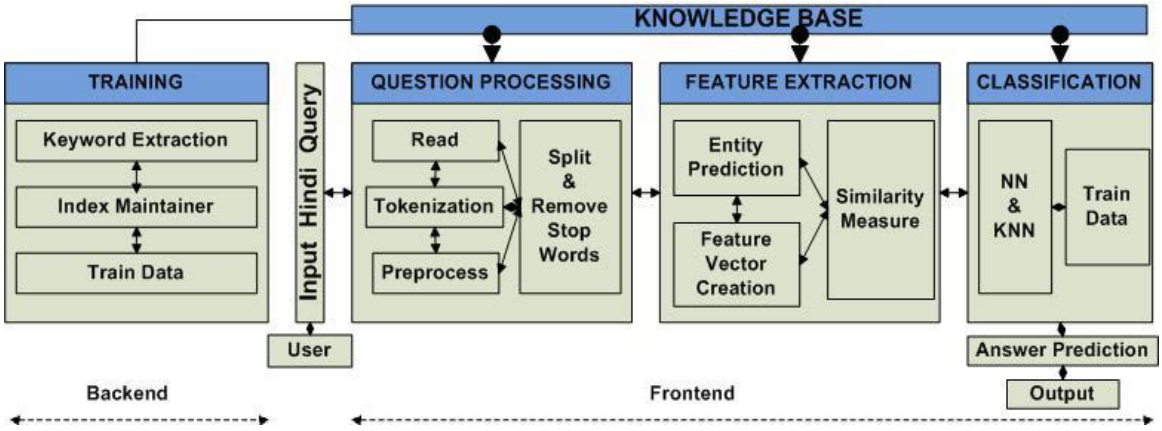

Figure 1. Architecture of QA system

3.2.2Feature Extraction

The second step of this module is to apply the feature extraction method on token words in which further two functions are performed Entity prediction and Feature Vector creation. Entity Prediction is the step to evaluate of performance and for this different similarity measurement methods can be used to measure the similarity between the two texts. The developed system has compared results using Cosine, Dice, Euclidean distance, Jaccard, JaroWinkler, SmithWaterman, Levenshtein, MongeElkan and NeedlemanWunch.The best result has been obtained from SmithWaterman. Finally, Feature Vector creation checks what type of question is asked by the user and related to that type of question of the index value is found from the knowledge base.

3.2.3 Classification method

In the final step of the frontend, the classification method is applied on train dataset by using the NN or KNN algorithm. NN algorithm was the first algorithm used to resolve the solution to TSP (Travelling Salesman Problem).NN executes quickly & easy to implement. KNN algorithm is a type of supervised learning algorithm

[2].That has been used in many applications like pattern recognition, data mining, image recognition etc. KNN algorithm is used for distance computing and distance ranking [10]. The formula for distance calculation using KNN is:

$$d(A_i, B_i) = \sqrt{\sum_{r=1}^{n} \left(x_r(A_i) - x_r(A_j)\right)^2} \quad (1)$$

Each distance enumeration method is self-determined and this distance calculation method helps in finding the K nearest neighbours for the each query. This algorithm is categorized on the basis of prior training data. After calculating the distance, it sorts the distances from the training data and finds the minimum distance and takes the maximum nearest neighbours.

3.2.4 Knowledge Base initialization

All frontend and backend working are performed by using the KB (Knowledge base) that takes training data as the input. The output answer is provided by the system after performing answer prediction method using the KB.

4. Implementation of Hindi QA System

In this web based QA system in which user entered a question in Hindi language and tokenized the query and designed a user interface system by using Net Beans IDE 7.3 on VMware workstation version 12.0[11]. In which Autocomplete method used that helps to find out the type of question which is entered by the user. Related to that type question all questions are shown in a given list and user can pick the required one from the list. Shown in fig.2 (a).

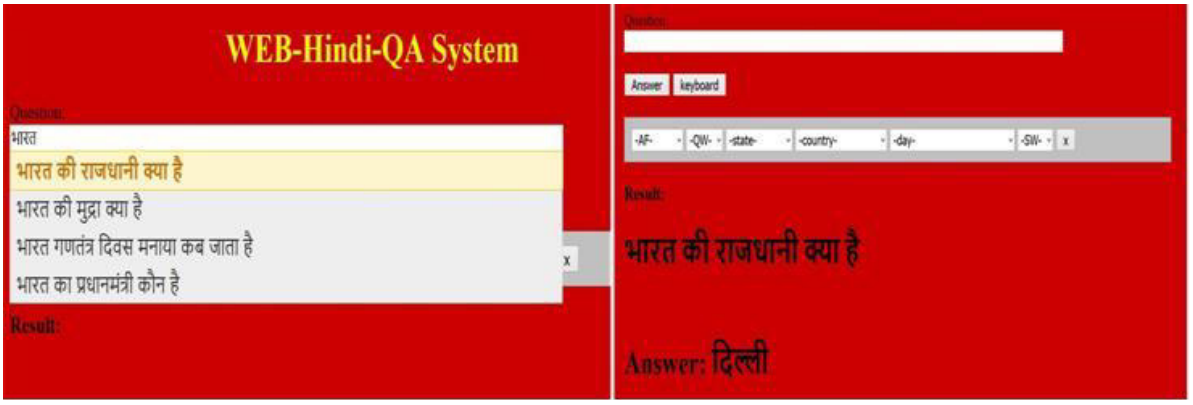

Fig2.(a) User interface of QA system(b). Result

In a web based QA system a keybord in which (QF,AF,QW) three fields are shown and user can enter a question with the help of keybord and autopartion method helps to find out what type of query asked by the user.for ex user select a words from the keybord 'भारत', 'राजधानी', 'क्या' as shown in fig.2(b) and then click on "Answer" button we get the result i.e 'दिल्ली'.

5. Testing and Results

5.1 Comparison Of Different Similarity Functions:

By using the similarity functions two different but similar texts have been compared and then the corresponding values have been generated. The contexts words which are extracted from the questions are compared. The priority is given to that context word which is used many times. Sometimes the user may type improper query, incomplete question or may do some spelling mistakes. The improper query is matched from the list of the database and return the similar values.

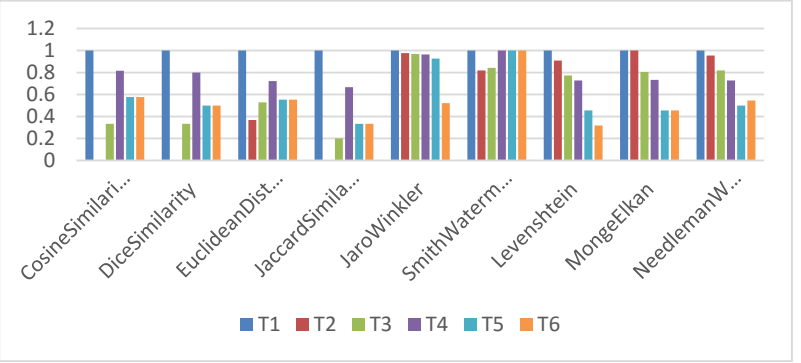

Fig .3 Comparison of different Similarity Functions

The system uses five text combination: first, with exact same text "केन्द्रीयउत्पादशुल्क" and second, with hyphon "केन्द्रीय-उत्पाद-शुल्क" and third with spelling mistake "केन्द्रीयउपादशुक" fourth with missing character "केन्द्रीयउत्पाद " and finally in fifth and sixth a single charcter has been taken i.e. "केन्द्रीय" , "स्वतंत्रता" ". The output values are as shown in Table1 and figure 3.Sometimes two meanings arise from the same context word but the system does not predict which the exact meaning of that question is and then user has to check these context words from the trained data set. Each function returns a different values corresponding to their context. The value of which function is greater is used as result. In this whole process a threshold value is set and match this value with every function. The value of which function, approximate or equal to threshold value, is taken. On testing, conclude that the best similarity measure is Smith Waterman for both misspelled words and multi-phases word with their actual word.

Table 1. Comparison of Different Similarity Function

केन्द्रीयउत्पादशुल्क	केन्द्रीयउत्पादशुल्क	केन्द्रीय-उत्पाद-शुल्क	केन्द्रीयउपादशुक	केन्द्रीयउत्पाद	केन्द्रीय	स्वतंत्रता
	T1	T2	T3	T4	T5	T6
CosineSimilarity	1	0	0.33333334	0.81649655	0.57735026	0.57735026
DiceSimilarity	1	0	0.33333334	0.8	0.5	0.5
EuclideanDistance	1	0.36754447	0.52859545	0.7226499	0.5527864	0.5527864
JaccardSimilarity	1	0	0.2	0.6666667	0.33333334	0.33333334
JaroWinkler	1	0.9757576	0.96874	0.9636364	0.92727274	0.521645
SmithWaterman	1	0.8181818	0.84210527	1	1	1
Levenshtein	1	0.9090909	0.77272725	0.72727275	0.45454544	0.3181818
MongeElkan	1	1	0.8055556	0.73333335	0.45555556	0.45555556
NeedlemanWunch	1	0.95454544	0.8181818	0.72727275	0.5	0.5454545

5.2 Comparison Of Different Data Set:

The two different data set are compared by using two different classification algorithms NN and K-NN. NN algorithm dispatches to find the closest point of the query from the given data set. NN measures the distance using K=1. K-NN measures the maximum nearest neighbours.KNN gives the better results than by NN algorithm. This comaprision performed with the help of two different data sets. As shown in fig.4for Testset-1 KNN gives 97%& NN gives 93.33%& for Testset-2 KNN gives 95.6% & NN gives 95% correct result.

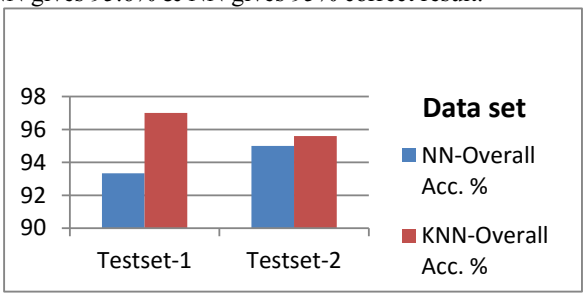

Fig .4: Comparison using differnt data sets

6. Future work and Conclusion

A system called Hindi web QA system that is capable to provide the exact answer of the question asked by the user has been developed. The system uses classification method which is able to work on a large data set and this data set is trained manually. The future work can be to make system multi-lingual in which user can put the question in any language and also, size of data set can also be increased**.**

References

1. Khalid MA, Jijkoun V and Rijke MD :Machine Learning for Question Answering from Tabular Data, 18th International Workshop on Database and Expert Systems Applications, pp. 392-396, 2007, IEEE.
2. Gharehchopogh FS and Lotfi Y : Machine Learning based Question Classification Methods in the Question Answering Systems, International Journal of Innovation and Applied Studies,Vol. 4 No. 2, pp. 264-273, Oct. 2013.
3. Sahu S, Vasnik N and Roy D : Prashnottar: A Hindi Question Answering System, International Journal of Computer Science & Information Technology (IJCSIT) Vol 4, No 2, pp. 149-158, April 2012.
4. Khillare SA, Shelke BA, Mahender CA : Comparative Study on Question Answering Systems and Techniques , 2014, IJARCSSE.
5. Dwivedi SK, Singh V : Research and reviews in question answering system, International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA), pp. 417-424, 2013.
6. Hirschman L,Gaizauskas R : Natural language question answering: the view from here, Vol. 7, pp. 275-300, 2001. Cambridge University Press.
7. Stalin S, Pandey R, Barskar R : Web Based Application for Hindi Question Answering System, International Journal of Electronics and Computer Science Engineering, Vol. 2, pp. 72-78, 2012.
8. Loni B : A Survey of State-of-the-Art Methods on Question Classification, 2011.
9. Bagde S, Dua M ,Virk ZS :Comparison of Different Similarity Functions on Hindi QA System, 2014 under publication.
10. Yan Z, Xu C: Combining KNN Algorithm and Other Classifiers, pp. 800-805, 2010, IEEE.
11. Kumar R, Dua M, Jindal S : D-HIRD: Domain-Independent Hindi Language Interface to Relational Database , pp. 81-86, 2014, IEEE.